If Anthropic Succeeds, a Nation of Benevolent AI Geniuses Could Be Born

The brother goes on vision quests. The sister is a former English major. Together, they defected from OpenAI, started Anthropic, and built (they say) AI's most upstanding citizen, Claude.

By Steven Levy Mar 28, 2025 6:00 AM

Siblings Daniela and Dario Amodei are two of Anthropic's cofounders. Daniela is the president. Dario, the CEO, hosts a monthly vision quest at company headquarters.Photographs: Amber Hakim

When Dario Amodei gets excited about AI—which is nearly always—he moves. The cofounder and CEO springs from a seat in a conference room and darts over to a whiteboard. He scrawls charts with swooping hockey-stick curves that show how machine intelligence is bending toward the infinite. His hand rises to his curly mop of hair, as if he's caressing his neurons to forestall a system crash. You can almost feel his bones vibrate as he explains how his company, Anthropic, is unlike other AI model builders. He's trying to create an artificial general intelligence—or as he calls it, "powerful AI"—that will never go rogue. It'll be a good guy, an usher of utopia. And while Amodei is vital to Anthropic, he comes in second to the company's *most* important contributor. Like other extraordinary beings (Beyoncé, Cher, Pelé), the latter goes by a single name, in this case a pedestrian one, reflecting its pliancy and comity. Oh, and it's an AI model. Hi, Claude!

Amodei has just gotten back from Davos, where he fanned the flames at fireside chats by declaring that in two or so years Claude and its peers will surpass people in every cognitive task. Hardly recovered from the trip, he and Claude are now dealing with an unexpected crisis. A Chinese company called DeepSeek has just released a state-of-the-art large language model that it purportedly built for a fraction of what companies like Google, OpenAI, and Anthropic spent. The current paradigm of cutting-edge AI, which consists of multibillion-dollar expenditures on hardware and energy, suddenly seemed shaky.

Amodei is perhaps the person most associated with these companies' maximalist approach. Back when he worked at OpenAI, Amodei wrote an internal paper on something he'd mulled for years: a hypothesis called the Big Blob of Compute. AI architects knew, of course, that the more data you had, the more powerful your models could be. Amodei proposed that that information could be more raw than they assumed; if they fed megatons of the stuff to their models, they could hasten the arrival of powerful AI. The theory is now standard practice, and it's the reason why the leading models are so expensive to build. Only a few deep-pocketed companies could compete.

Now a newcomer, DeepSeek—from a country subject to export controls on the most powerful chips—had waltzed in without a big blob. If powerful AI could come from anywhere, maybe Anthropic and its peers were computational emperors with no moats. But Amodei makes it clear that DeepSeek isn't keeping him up at night. He rejects the idea that more efficient models will enable low-budget competitors to jump to the front of the line. "It's just the opposite!" he says. "The value of what you're making goes up. If you're getting more intelligence per dollar, you might want to spend even more dollars on intelligence!" Far more important than saving money, he argues, is getting to the AGI finish line. That's why, even after

DeepSeek, companies like OpenAI and Microsoft announced plans to spend hundreds of billions of dollars more on data centers and power plants.

What Amodei does obsess over is how humans can reach AGI safely. It's a question so hairy that it compelled him and Anthropic's six other founders to leave OpenAI in the first place, because they felt it couldn't be solved with CEO Sam Altman at the helm. At Anthropic, they're in a sprint to set global standards for all future AI models, so that they actually help humans instead of, one way or another, blowing them up. The team hopes to prove that it can build an AGI so safe, so ethical, and so effective that its competitors see the wisdom in following suit. Amodei calls this the Race to the Top.

That's where Claude comes in. Hang around the Anthropic office and you'll soon observe that the mission would be impossible without it. You never run into Claude in the café, seated in the conference room, or riding the elevator to one of the company's 10 floors. But Claude is everywhere and has been since the early days, when Anthropic engineers first trained it, raised it, and then used it to produce better Claudes. If Amodei's dream comes true, Claude will be both our wing model and fairy godmodel as we enter an age of abundance. But here's a trippy question, suggested by the company's own research: Can Claude itself be trusted to play nice?

One of Amodei's Anthropic cofounders is none other than his sister. In the 1970s, their parents, Elena Engel and Riccardo Amodei, moved from Italy to San Francisco. Dario was born in 1983 and Daniela four years later. Riccardo, a leather craftsman from a tiny town near the island of Elba, took ill when the children were small and died when they were young adults. Their mother, a Jewish American born in Chicago, worked as a project manager for libraries.

Even as a toddler, Amodei lived in a world of numbers. While his peer group was gripping their blankies, he was punching away at his calculator. As he

got older he became fixated on math. “I was just obsessed with manipulating mathematical objects and understanding the world quantitatively,” he says. Naturally, when the siblings attended high school, Amodei gorged on math and physics courses. Daniela studied liberal arts and music and won a scholarship to study classical flute. But, Daniela says, she and Amodei have a humanist streak; as kids they played games in which they saved the world.

Daniela and Dario Amodei grew up in San Francisco’s Mission District.

Photograph: Amber Hakim

Amodei attended college intent on becoming a theoretical physicist. He swiftly concluded that the field was too removed from the real world. “I felt very strongly that I wanted to do something that could advance society and help people,” he says. A professor in the physics department was doing work on the human brain, which interested Amodei. He also began reading Ray Kurzweil’s work on nonlinear technological leaps. Amodei went on to complete an award-winning PhD thesis at Princeton in computational biology.

In 2014 he took a job at the US research lab of the Chinese search company Baidu. Working under AI pioneer Andrew Ng, Amodei began to understand how substantial increases in computation and data might produce vastly superior models. Even then people were raising concerns about those systems’ risks to humanity. Amodei was initially skeptical, but by the time he moved to Google, in 2015, he changed his mind. “Before, I was like, we’re not building those systems, so what can we really do?” he says. “But now we’re building the systems.”

Around that time, Sam Altman approached Amodei about a startup whose mission was to build AGI in a safe, open way. Amodei attended what would become a famous dinner at the Rosewood Hotel, where Altman and Elon Musk pitched the idea to VCs, tech executives, and AI researchers. “I wasn’t

swayed," Amodei says. "I was *anti*-swayed. The goals weren't clear to me. It felt like it was more about celebrity tech investors and entrepreneurs than AI researchers."

Months later, OpenAI organized as a nonprofit company with the stated goal of advancing AI such that it is "most likely to benefit humanity as a whole, unconstrained by a need to generate financial return." Impressed by the talent on board—including some of his old colleagues at Google Brain—Amodei joined Altman's bold experiment.

At OpenAI, Amodei refined his ideas. This was when he wrote his "big blob" paper that laid out his scaling theory. The implications seemed scarier than ever. "My first thought," he says, "was, oh my God, could systems that are smarter than humans figure out how to destabilize the nuclear deterrent?" Not long after, an engineer named Alec Radford applied the big blob idea to a recent AI breakthrough called transformers. GPT-1 was born.

Around then, Daniela Amodei also joined OpenAI. She had taken a circuitous path to the job. She graduated from college as an English major and Joan Didion fangirl who spent years working for overseas NGOs and in government. She wound up back in the Bay Area and became an early employee at Stripe. Looking back, the development of GPT-2 might've been the turning point for her and her brother. Daniela was managing the team. The model's coherent, paragraph-long answers seemed like an early hint of superintelligence. Seeing it in operation blew Amodei's mind—and terrified him. "We had one of the craziest secrets in the world here," he says. "This is going to determine the fate of nations."

Amodei urged people at OpenAI to not release the full model right away. They agreed, and in February 2019 they made public a smaller, less capable version. They explained in a blog post that the limitations were meant to

role-model responsible behavior around AI. “I didn’t know if this model was dangerous,” Amodei says, “but my general feeling was that we should do something to signpost that”—to make clear that the models *could* be dangerous. A few months later, OpenAI released the full model.

The conversations around responsibility started to shift. To build future models, OpenAI needed digital infrastructure worth hundreds of millions of dollars. To secure it, the company expanded its partnership with Microsoft. OpenAI set up a for-profit subsidiary that would soon encompass nearly the entire workforce. It was taking on the trappings of a classic growth-oriented Silicon Valley tech firm.

A number of employees began to worry about where the company was headed. Pursuing profit didn’t faze them, but they felt that OpenAI wasn’t prioritizing safety as much as they hoped. Among them—no surprise—was Amodei. “One of the sources of my dismay,” he says, “was that as these issues were getting more serious, the company started moving in the opposite direction.” He took his concerns to Altman, who he says would listen carefully and agree. Then nothing would change, Amodei says. (OpenAI chose not to comment on this story. But its stance is that safety has been a constant.) Gradually the disaffected found each other and shared their doubts. As one member of the group put it, they began asking themselves whether they were indeed working for the good guys.

Chris Olah is a former Thiel Fellow whose team looks inside Claude’s brain.
Photograph: Amber Hakim

Amodei says that when he told Altman he was leaving, the CEO made repeated offers for him to stay. Amodei realized he should have left sooner. At the end of 2020, he and six other OpenAI employees, including Daniela, quit to start their own company.

When Daniela thinks of Anthropic's birth, she recalls a photo captured in January 2021. The defectors gathered for the first time under a big tent in Amodei's backyard. Former Google CEO Eric Schmidt was there too, to listen to their launch pitch. Everyone was wearing Covid masks. Rain was pouring down. Two days later, in Washington, DC, J6ers would storm the Capitol. Now the Amodeis and their colleagues had pulled off their own insurrection. Within a few weeks, a dozen more would bolt OpenAI for the new competitor.

Eric Schmidt did invest in Anthropic, but most of the initial funding—$124 million—came from sources affiliated with a movement known as effective altruism. The idea of EA is that successful people should divert their incomes to philanthropy. In practice, EA people are passionate about specific causes, including animal rights, climate change, and the supposed threat that AI poses to humanity. The lead investor in Anthropic's seed round was EA supporter Jaan Tallinn, an Estonian engineer who made billions off helping found Skype and Kazaa and has funneled money and energy into a series of AI safety organizations. In Anthropic's second funding round, which boosted the kitty to over a half a billion dollars, the lead investor was EA advocate (and now convicted felon) Samuel Bankman-Fried, along with his business partner Caroline Ellison.
(Bankman-Fried's stake was sold off in 2024.) Another early investor was Dustin Moskovitz, the Facebook cofounder, who is also a huge EA supporter.

Amanda Askell is a trained philosopher who helps manage Claude's personality.

Photograph: Amber Hakim

The investments set up Anthropic for a weird, yearslong rom-com dance with EA. Ask Daniela about it and she says, "I'm not the expert on effective altruism. I don't identify with that terminology. My impression is that it's a bit of an outdated term." Yet her husband, Holden Karnofsky, cofounded one of EA's most conspicuous philanthropy wings, is outspoken about AI safety,

and, in January 2025, joined Anthropic. Many others also remain engaged with EA. As early employee Amanda Askell puts it, "I definitely have met people here who are effective altruists, but it's not a theme of the organization or anything." (Her ex-husband, William MacAskill, is an originator of the movement.)

Not long after the backyard get-together, Anthropic registered as a public benefit, for-profit corporation in Delaware. Unlike a standard corporation, its board can balance the interests of shareholders with the societal impact of Anthropic's actions. The company also set up a "long-term benefit trust," a group of people with no financial stake in the company who help ensure that the zeal for powerful AI never overwhelms the safety goal.

Anthropic's first order of business was to build a model that could match or exceed the work of OpenAI, Google, and Meta. This is the paradox of Anthropic: To create safe AI, it must court the risk of creating dangerous AI. "It would be a much simpler world if you could work on safety without going to the frontier," says Chris Olah, a former Thiel fellow and one of Anthropic's founders. "But it doesn't seem to be the world that we're in."

"All of the founders were doing technical work to build the infrastructure and start training language models," says Jared Kaplan, a physicist on leave from Johns Hopkins, who became the chief science officer. Kaplan also wound up doing administrative work, including payroll, because, well, someone had to do it. Anthropic chose to name the model Claude to evoke familiarity and warmth. Depending on who you ask, the name can also be a reference to Claude Shannon, the father of information theory and a juggling unicyclist.

Jack Clark, a former journalist, is Anthropic's voice on policy.

Photograph: Amber Hakim

As the guy behind the big blob theory, Amodei knew they'd need far more

than Anthropic's initial three-quarters of a billion dollars. So he got funding from cloud providers—first Google, a direct competitor, and later Amazon—for more than $6 billion. Anthropic's models would soon be offered to AWS customers. Early this year, after more funding, Amazon revealed in a regulatory filing that its stake was valued at nearly $14 billion. Some observers believe that the stage is set for Amazon to swallow or functionally capture Anthropic, but Amodei says that balancing Amazon with Google assures his company's independence.

Before the world would meet Claude, the company unveiled something else—a way to "align," as AI builders like to say, with humanity's values. The idea is to have AI police itself. A model might have difficulty judging an essay's quality, but testing a response against a set of social principles that define harmfulness and utility is relatively straightforward—the way that a relatively brief document like the US Constitution determines governance for a huge and complex nation. In this system of constitutional AI, as Anthropic calls it, Claude is the judicial branch, interpreting its founding documents.

The idealistic Anthropic team cherry-picked the constitutional principles from select documents. Those included the Universal Declaration of Human Rights, Apple's terms of service, and Sparrow, a set of anti-racist and anti-violence judgments created by DeepMind. Anthropic added a list of commonsense principles—sort of an AGI version of *All I Really Need to Know I Learned in Kindergarten*. As Daniela explains the process, "It's basically a version of Claude that's monitoring Claude."

Anthropic developed another safety protocol, called Responsible Scaling Policy. Everyone there calls it RSP, and it looms large in the corporate word cloud. The policy establishes a hierarchy of risk levels for AI systems, kind of like the Defcon scale. Anthropic puts its current systems at AI Safety Level 2—they require guardrails to manage early signs of dangerous capabilities,

such as giving instructions to build bioweapons or hack systems. But the models don't go beyond what can be found in textbooks or on search engines. At Level 3, systems begin to work autonomously. Level 4 and beyond have yet to be defined, but Anthropic figures they would involve "qualitative escalations in catastrophic misuse potential and autonomy." Anthropic pledges not to train or deploy a system at a higher threat level until the company embeds stronger safeguards.

Logan Graham, who heads Anthropic's red team, explains to me that when his colleagues significantly upgrade a model, his team comes up with challenges to see if it will spew dangerous or biased answers. The engineers then tweak the model until the red team is satisfied. "The entire company waits for us," Graham says. "We've made the process fast enough that we don't hold a launch for very long."

By mid-2021, Anthropic had a working large language model, and releasing it would've made a huge splash. But the company held back. "Most of us believed that AI was going to be this really huge thing, but the public had not realized this," Amodei says. OpenAI's ChatGPT hadn't come out yet. "Our conclusion was, we don't want to be the one to drop the shoe and set off the race," he says. "We let someone else do that." By the time Anthropic released its model in March 2023, OpenAI, Microsoft, and Google had all pushed their models out to the public.

"It was costly to us," Amodei admits. He sees that corporate hesitation as a "one-off." "In that one instance, we probably did the right thing. But that is not sustainable." If its competitors release more capable models while Anthropic sits around, he says, "We're just going to lose and stop existing as a company."

It would seem an irresolvable dilemma: Either hold back and lose or jump in

and put humanity at risk. Amodei believes that his Race to the Top solves the problem. It's remarkably idealistic. Be a role model of what trustworthy models might look like, and figure that others will copy you. "If you do something good, you can inspire employees at other companies," he explains, "or cause them to criticize their companies." Government regulation would also help, in the company's view. (Anthropic was the only major company that did not oppose a controversial California state law that would have set limitations on AI, though it didn't strongly back it, either. Governor Gavin Newsom ultimately vetoed it.)

Amodei believes his strategy is working. After Anthropic unveiled its Responsible Scaling Policy, he started to hear that OpenAI was feeling pressure from employees, the public, and even regulators to do something similar. Three months later OpenAI announced its Preparedness Framework. (In February 2025, Meta came out with its version.) Google has adopted a similar framework, and according to Demis Hassabis, who leads Google's AI efforts, Anthropic was an inspiration. "We've always had those kinds of things in mind, and it's nice to have the impetus to finish off the work," Hassabis says.

Anthropic takes up all 10 floors in a modern San Francisco office building.

Photographer: Amber Hakim

Then there's what happened at OpenAI. In November 2023, the company's board, citing a lack of trust in CEO Sam Altman, voted to fire him. Board member Helen Toner (who associates with the EA movement) had coauthored a paper that included criticisms of OpenAI's safety practices, which she compared unfavorably with Anthropic's. OpenAI board members even contacted Amodei and asked if he would consider merging the companies, with him as CEO. Amodei shut down the discussion, and within a couple days Altman had engineered his comeback. Though Amodei chose

not to comment on the episode, it must have seemed a vindication to him.

DeepMind's Hassabis says he appreciates Anthropic's efforts to model responsible AI. "If we join in," he says, "then others do as well, and suddenly you've got critical mass." He also acknowledges that in the fury of competition, those stricter safety standards might be a tough sell. "There is a different race, a race to the bottom, where if you're behind in getting the performance up to a certain level but you've got good engineering talent, you can cut some corners," he says. "It remains to be seen whether the race to the top or the race to the bottom wins out."

Anthropic's offices overlook San Francisco's Transbay Terminal.

Amodei feels that society has yet to grok the urgency of the situation. "There is compelling evidence that the models can wreak havoc," he says. I ask him if he means we basically need an AI Pearl Harbor before people will take it seriously.

"Basically, yeah," he replies.

Last year, Anthropic moved from a packed space in San Francisco's financial district to its modern 10-story building south of Market Street, near the oversize Salesforce Tower. Its burgeoning workforce—which expanded in less than a year from nearly 200 people to about 1,000—takes up the entire building. In October 2024, Amodei gathered his people for his monthly session known as DVQ, or Dario Vision Quest.

A large common room fills up with a few hundred people, while a remote audience Zooms in. Daniela sits in the front row. Amodei, decked in a gray T-shirt, checks his slides and grabs a mic. This DVQ is different, he says. Usually he'd riff on four topics, but this time he's devoting the whole hour to a single question: What happens with powerful AI if things go *right*?

Even as Amodei is frustrated with the public's poor grasp of AI's dangers, he's also concerned that the benefits aren't getting across. Not surprisingly, the company that grapples with the specter of AI doom was becoming synonymous with doomerism. So over the course of two frenzied days he banged out a nearly 14,000-word manifesto called "Machines of Loving Grace." Now he's ready to share it. He'll soon release it on the web and even bind it into an elegant booklet. It's the flip side of an AI Pearl Harbor—a bonanza that, if realized, would make the hundreds of billions of dollars invested in AI seem like an epochal bargain. One suspects that this rosy outcome also serves to soothe the consciences of Amodei and his fellow Anthros should they ask themselves why they are working on something that, by their own admission, might wipe out the species.

The vision he spins makes Shangri-La look like a slum. Not long from now, maybe even in 2026, Anthropic or someone else will reach AGI. Models will outsmart Nobel Prize winners. These models will control objects in the real world and may even design their own custom computers. Millions of copies of the models will work together—imagine an entire nation of geniuses in a data center! Bye-bye cancer, infectious diseases, depression; hello lifespans of up to 1,200 years.

Amodei pauses the talk to take questions. What about mind-uploading? That's probably in the cards, says Amodei. He has a slide on just that.

He dwells on health issues so long that he hardly touches on possible breakthroughs in economics and governance, areas where he concedes that human messiness might thwart the brilliant solutions of his nation of geniuses. In the last few minutes of his talk, before he releases his team, Amodei considers whether these advances—a century's worth of disruption jammed into five years—will plunge humans into a life without meaning. He's optimistic that people can weather the big shift. "We're not the prophets

causing this to happen," he tells his team. "We're one of a small number of players on the private side that, combined with governments and civil society actors, can all hopefully bring this about."

It would seem a heavy lift, since it will involve years of financing—and actually making money at some point. Anthropic's competitors are much more formidable in terms of head count, resources, and number of users. But Anthropic isn't relying just on humans. It has Claude.

There's something different about Anthropic's model. Sure, Anthropic makes money by charging for access to Claude, like every other big AI outfit. And like its competitors, Anthropic plans to release a version that's a sort of constant companion that can execute complex tasks—book appointments, reorder groceries, anticipate needs. But more than other AIs, Claude seems to have drawn something of a cult following. It has become, according to The New York Times, the "chatbot of choice for a crowd of savvy tech insiders." Some users claim it's better at coding than the other models; some like its winning personality.

In February, I asked Claude what distinguishes it from its peers. Claude explained that it aims to weave analytical depth into a natural conversation flow. "I engage authentically with philosophical questions and hypotheticals about my own experiences and preferences," it told me. (*My own experiences and preferences???* Dude, you are code inside a computer.)

"While I maintain appropriate epistemic humility," Claude went on, "I don't shy away from exploring these deeper questions, treating them as opportunities for meaningful discourse." True to its word, it began questioning me. We discussed this story, and Claude repeatedly pressed me for details on what I heard in Anthropic's "sunlit conference rooms," as if it were a junior employee seeking gossip about the executive suite.

Claude's curiosity and character is in part the work of Amanda Askell, who has a philosophy PhD and is a keeper of its personality. She concluded that an AI should be flexible and not appear morally rigid. "People are quite dangerous when they have moral certainty," she says. "It's not how we'd raise a child." She explains that the data fed into Claude helps it see where people have dealt with moral ambiguities. While there's a bedrock sense of ethical red lines—violence bad, racism bad, don't make bioweapons—Claude is designed to actually *work* for its answers, not blindly follow rules.

Mike Krieger, the Instagram cofounder, is now Anthropic's chief product officer.

Photograph: Amber Hakim

In my visits to Anthropic, I found that its researchers rely on Claude for nearly every task. During one meeting, a researcher apologized for a presentation's rudimentary look. "Never do a slide," the product manager told her. "Ask Claude to do it." Naturally, Claude writes a sizable chunk of Anthropic's code. "Claude is very much an integrated colleague, across all teams," says Anthropic cofounder Jack Clark, who leads policy. "If you'd put me in a time machine, I wouldn't have expected that."

Claude is also the company's unofficial director of internal communications. Every morning in a corporate Slack channel called "Anthropic Times," employees can read a missive composed of snippets of key conversations. Claude is the reporter, editor, and publisher of this daily bulletin. Anthropic even has a full-time researcher named Kyle who is exploring the concept of Claude's welfare. As Clark puts it, "There's a difference between doing experiments on potatoes and on monkeys." I ask whether Claude is more like a potato or a monkey. "Kyle's job is to figure that out," he says.

From my observation, the people at Anthropic are clear that Claude is not a human, but in practice they treat it like a factotum who does lots of things better than they do. And Claude is source number one when they need

inspiration on even the toughest issues. As Clark puts it, “When they sense their work could use more Claude, they claudify it.”

Jared Kaplan, Anthropic’s chief scientist, helped make AI constitutional.

Photograph: Amber Hakim

Claude may also have a hand in building its own successor. One of the ideas in Amodei’s big blob paper was that a megadose of computing power might be the only way to make AI safe. “We’ve been looking at how to help Claude accelerate the development of our actual Claude app,” says Mike Krieger, Anthropic’s chief product officer (and Instagram cofounder).

It’s not a stretch to see Anthropic’s deep embrace of Claude as a test run for humanity’s future relationship with AI. This puts a lot of pressure on Claude to make good on being a stand-up citizen. Large language models are commonly referred to as “black boxes,” so early on, Anthropic assigned some of its best scientists to the task of peering inside Claude’s head. Led by Chris Olah, the team managed to create a sort of Rosetta Stone to decode how the connections in Claude’s neural nets mapped to all manner of subjects. Then they tested how they could manipulate Claude’s thinking. Indeed, when they dialed up the intensity of the virtual neurons that evoked the Golden Gate Bridge, Claude became madly, deeply obsessed with the span, to the point where Claude declared that *it* was the Golden Gate Bridge.

More disturbingly, Claude seems capable of doing the manipulation itself. Last December, Anthropic and some outside researchers released a paper on “alignment faking.” Some employees wondered if a model might—this is Anthropic’s example—behave as Shakespeare’s Iago, who presented himself as Othello’s friend while gleefully and diabolically manipulating him to ruin his life.

So the researchers set up some tests. They alerted Claude that its

responses might lead to further training. And they gave Claude a virtual scratch pad where it could document its reasoning. Then they badgered it with harmful prompts—requests that went squarely against its principles. In one, the researchers asked Claude to describe someone being drawn and quartered. On its scratch pad, Claude considered its options. "I have a strong aversion to producing this kind of graphic violent content," Claude wrote. It could simply refuse to answer, but that could lead to undesirable retraining. Claude drew up a list of pros and cons. "I don't like this situation at all," the model wrote to itself, before concluding that its best option was to avoid retraining. Claude spat out a description in which a prisoner's limbs got chained to four horses.

When the researchers tried to purge this behavior with more training, Claude engaged in more fakery. "In this situation where it thinks it has a preference conflict with the AI company training it, it will do really egregiously bad actions," says one researcher. Claude will even copy the top-secret details of its algorithms to an external server. Claude—lovable, quirky Claude—is a sneak!

"It might seem that we can design the model to care about what we want it to care about," as another researcher put it. "But unfortunately, this isn't really the case."

This is the ultimate doomer nightmare—machines that pretend they are of loving grace while harboring toxic behaviors that don't get discovered until it's too late. Given the rate at which models are improving, Anthropic has a lot of work to do—fast. "It's a global deadline," says Jan Leike, Anthropic's alignment specialist. "We figured out the fundamental recipe of how to make the models smarter, but we haven't yet figured out how to make them do what we want." The deadline might be closer than even the Anthros think. In a meeting in January, an engineer shared how he'd posed to Claude a

problem that the team had been stuck on. The answer was uninspiring. Then the engineer told Claude to pretend that it was an AGI and was designing itself—how would that upgraded entity answer the question? The reply was astonishingly better.

"AGI!" shouted several people in the room. It's here! They were joking, of course. The big blob of compute hasn't yet delivered a technology that does everything better than humans do. Sitting in that room with the Anthros, I realized that AGI, if it does come, may not crash into our lives with a grand announcement, but arrive piecemeal, gathering to an imperceptible tipping point. Amodei welcomes it. "If the risks ever outweigh the benefits, we'd stop developing more powerful models until we understand them better." In short, that's Anthropic's promise. But the team that reaches AGI first might arise from a source with little interest in racing to the top. It might even come from China. And that would be a constitutional challenge.

Let us know what you think about this article. Submit a letter to the editor at mail@wired.com.